Original Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

A SUPPORT VECTOR MACHINE BASED APPROACH FOR PREDICTION AND CLASSIFICATION OF POLYANION BINDING PROTEINS

M.UDAYAKUMAR*[1], R.SENTHILKUMAR[1], A.D.SHRIVATHSAN[2]

[1]School of Chemical and Biotechnology, SASTRA University, India
[2]School of Computing, SASTRA University, India

ABSTRACT

Polyanionic proteins are extremely abundant in both extracellular and intracellular environment of the cell. These binding proteins contain multiple positively charged regions which are involved in phosphorylation processes that play a vital role in interaction with cellular proteins and also involved in process of protein folding. The various classes of polyanion binding proteins used are actin, heparin, heparin sulfate and tubulin binding proteins. In this article, we are interested in protein sequence based classification of polyanion-binding and non-polyanion binding proteins. Firstly, protein sequence features like amino acid composition, dipeptide composition, hydrophobicity and hybrid combinations of these features were used to develop SVM modules. Then training and testing cycle were performed using the SVM^{light} software. These modules were then evaluated using the 10 fold cross-validation technique. Furthermore, the method was able to predict major classes of binding proteins based on amino acid composition(AAC), dipeptide composition (DPC), hydrophobicity (Hydro), (AAC + DPC), (AAC + Hydro) and (DPC + Hydro) with an accuracy of 68.2012%, 70.2796%, 53.8530%, 76.6861%, 64.3378%, and 71.5340% and it was also able to predict major subclasses of polyanion binding proteins using AAC, DPC , Hydro ,AAC + DPC,AAC + Hydro and DPC + Hydro with a maximum accuracy (92.80%,,94.44%,93.30%,93.29%,93.40%,93.33%),(57.69%,57.94%,57.32%,74.67%,59.73%,57.16%),(84.61%,88.88%,85.78%,86.67%,86.67 %,86.67%) (80.76%,83.22%,82.61% ,82.31%, 77.73%,78.44%) for heparan sulfate, actin , heparin and tubulin respectively. We obtained a good classification performance for the SVM classifier trained with combined feature of amino acid and dipeptide features.

KEYWORDS: **Polyanion-binding proteins, SVM, kernel function, UniParc, Heparin sulfate**

* **M.UDAYAKUMAR**
School of Chemical and Biotechnology, SASTRA University, India

Received on : 27-09-2016
Revised and Accepted on : 14-11-2016
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.1.b126-131

INTRODUCTION

The macromolecular structures are made up of basic molecular units like proteins and nucleic acids which are involved in functional cellular processes such as transcription, replication and recombination.[1] Many cellular macromolecules and macromolecular complexes (proteoglycans, DNA, RNA, ribosomes, actin microfilaments and microtubules) are polyanionic in nature and they interact with proteins and regulate different physiological and pathological functions.[2] These polyanion binding proteins have been classified into four categories as actin, heparin, heparin sulfate and tubulin.[3] These four classes of polyanions are called cellular polyanions. Brain tissues of neurodegenerative disease like Alzheimer's and Parkinson's have revealed that the presence of abnormal deposits of polyanion proteins.[4] There is a need of computational methods for the prediction of ployanion-binding proteinssuch as SVM[5] which is used in a variety of biological applications and also employed in various classification and regression tasks. SVM classifiers helps to classify the data into various categories using the internal data structures. The kernel function used here is the radial basis function (RBF) is used to classify the data, using one versus rest strategy. SVM^{light} is an implementation of Vapnik's Support Vector Machine which is used in text classification[6,7,] image recognition[8], microarray gene expression data analysis[9] and protein fold recognition.[10] SVM^{light} is a freely downloadable software package from the world wide web.[11]

MATERIALS AND METHODS

Datasets

The implementation of SVM classifier for Polyanion binding proteins is shown in Figure1. The polyanion-binding proteins were downloaded from UniParc database. UniParc is the comprehensive non-redundant protein sequence database.[12] We removed the redundancy of the downloaded sequences using the program CDHIT.[13] After the removal of redundancy, our final dataset used in this study contains 7827 polyanion-binding proteins and 941 non-polyanion binding proteins. These 7827 polyanion-binding proteins were further classified into four different classes, such as 4939 actin binding proteins, 900 heparin binding proteins, 545 heparan sulfate binding proteins and 1443 tubulin binding proteins. Actin involves in protein-protein interaction whereas Tubulin plays a multitude roles in cell structure transport. A wide variety of vital functions are played by Heparan sulfate (HS) proteoglycans, in many biological process in the animal kingdom.[14] Heparin remains the most widely used parentral antithrombotic and has the ability to inhibit molecular interactions.[15,16] About 75% of the randomly chosen sequences from each class were considered as the training dataset. Similarly, 25% of the randomly chosen sequences from each class were considered as the testing dataset which is shown in Table1.

Table 1
Polyanion protein Sequence Dataset used in this study

Total Number of proteins used for classification			
Class	Total number of proteins	Number of proteins for Training Set	Number of proteins for Testing Set
Actin	4939	3704	1235
Heparan Sulfate	545	408	137
Heparin	900	675	225
Tubulin	1443	1082	361
Non-Polyanion	941	705	236

Input features and Performance Evaluation

Each sequence in the dataset is represented in the form of 6 different features. The features used in the study were classified into two categories: normal features and the hybrid combination of features.The normal features included amino acid composition, dipeptide composition and hydrophobicity. For amino acid composition, a protein is represented by a vector of 20 dimensions, while for dipeptide composition, a protein is represented by 400 dimensions (20 X 20). Similarly, calculation of hydrophobicity for a protein contains 20 dimensions. To calculate the hydrophobicity, the hydropathy index values given by Kyte –Doolittle.[17]

Amino Acid Composition (AAC)

The following equation was used to calculate the AAC,

$$F_i = \frac{\text{Total number of amino acid i}}{\text{Total number of amino acids in the proteins}} \qquad (A)$$

Dipeptide Composition (DPC)

The following equation was used to calculate the DPC,

$$F_j = \frac{\text{Total number of dipeptide j}}{\text{Total number of possible dipeptides in the protein}} \qquad (B)$$

Hybrid combination of features

The hybrid combination of features includes the amino acid and dipeptide composition (hybrid 1), Amino acid composition and hydrophobicity (hybrid 2) and Dipeptide composition and hydrophobicity (hybrid 3). Hybrid 1 consisted of 420 dimensions (20 amino acid + 400 dipeptide), while hybrid 2 had 40 dimensions (20 amino acid + 20 hydrophobicity) and hybrid 3 had 420 dimensions (400 dipeptide + 20 hydrophobicity). The various features used in the study were calculated using Perl script.

SVM Multi-class Classification

This type of classification was performed using SVM^{light} software. Using recursive approach, the developed SVM modules were subjected to training and testing procedures on different combination of features (i.e hybrids). The one-versus-rest strategy was used to separate data into training and test samples.[18]. The RBF kernel function was used to classify the dataset. For training and classifying the dataset, two modules were used: svm_classify and svm_learn. The svm_classify module is used to train the dataset that contains all the classes in it. Similarly, the svm_learn module in SVM^{light} is used to predict the data. The parameters such as C and gamma values were optimized to obtain better results.

Performance of SVM standalone models

To have a reliable estimate, ten fold cross-validation method was used to evaluate the performance of the developed SVM model. To perform cross-validation technique, the dataset was randomly divided into 10 sets. Each set consists of all the classes of polyanion proteins in it. The process of training and testing was performed 10 times for each model. For each time, nine subsets are used as training data and the remaining one is used as test data. The overall performance of the model was calculated by considering the average performance over all the 10 sets. To obtain better results, work was done with the unbalanced dataset, the j and gamma parameters were used.[19] Gamma value is a parameter in radial basis kernel function. The accuracy values varied for each dataset (Table 1). The performance of the SVM was also measured using the following parameters.[20]

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100 \quad (C)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100 \quad (D)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \quad (E)$$

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)}} \quad (F)$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative. The Matthew's correlation coefficient ranges from –1 ≤MCC ≤1.

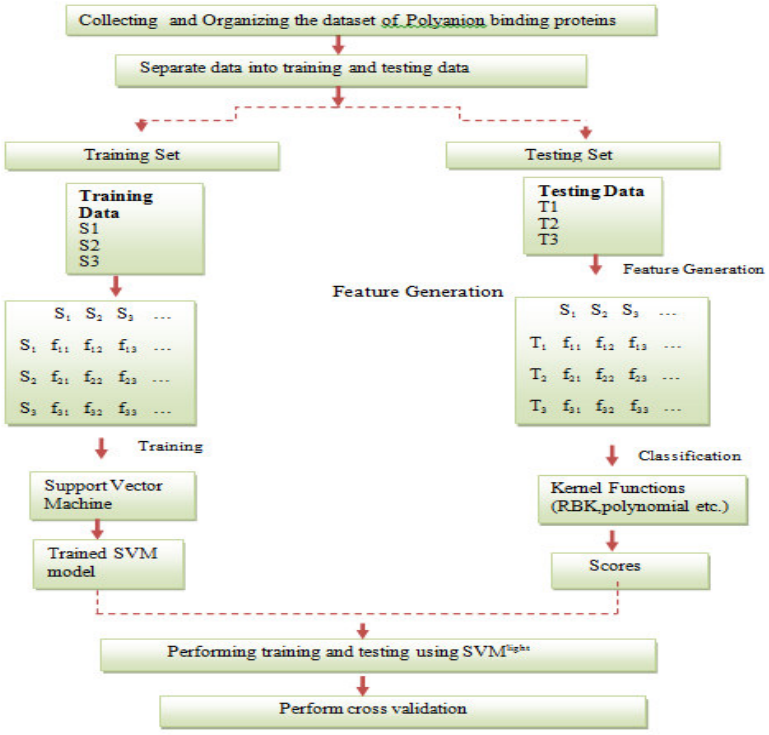

Figure 1
SVM model for prediction of Polyanion binding proteins

RESULTS AND DISCUSSIONS

The classifier worked with both polyanion-binding and non-polyanion binding proteins. The accuracy of values for the six features are tabulated in Table 2.The highest accuracy value is 76.68% with 83.99% specificity and 86.03% sensitivity for the hybrid combination of AAC and DPC and their respective MCC value is 0.67.This indicates that the best classification efficiency was achieved by the AAC and DPC when compared to other features.From Table 3, we obtained highest accuracy of 94.44% for the heparin sulfate class of polyanion binding proteins.

Table 2

Performance of various SVM modules: Accuracy and MCC values obtained for the various modules and their hybrid combinations.

Model	Feature	Dimension	Accuracy (in %)	MCC
Model1	AAC	20	68.2012	0.421
Model2	DPC	400	70.2796	0.363
Model3	Hydro	20	53.8530	0.1
Hybrid1	**AAC + DPC**	**420**	**76.6861**	**0.670**
Hybrid2	AAC + Hydro	40	64.3378	0.347
Hybrid3	DPC + Hydro	420	71.5340	0.307

AAC=Amino Acid Composition, DPC=Dipeptide Composition, Hydro=Hydrophobicity

Table 3

Performance of SVM modules of various classes (Actin,Heparan sulfate,Heparin, Tubulin) using the protein sequence features (AAC, DPC,Hydro)

Type of binding protein	Feature	Dimension	Accuracy (in %)
Actin	DPC	400	57.94
	Hydro	20	57.32
	AAC + DPC	420	74.67
	AAC + Hydro	40	59.73
	DPC + Hydro	420	57.16
Heparan sulfate	AAC	20	92.80
	DPC	**400**	**93.44**
	Hydro	20	93.30
	AAC + DPC	420	94.29
	AAC + Hydro	40	93.40
	DPC + Hydro	420	93.33
Heparin	AAC	20	84.61
	DPC	400	88.88
	Hydro	20	85.78
	AAC + DPC	420	86.67
	AAC + Hydro	40	86.67
	DPC + Hydro	420	86.67
Tubulin	AAC	20	80.76
	DPC	400	83.22
	Hydro	20	82.61
	AAC + DPC	420	82.31
	AAC + Hydro	40	77.73
	DPC + Hydro	420	78.44

AAC=Amino Acid Composition, DPC=Dipeptide Composition, Hydro=Hydrophobicity

Evaluation of SVM classifier

To further analyze the classifiers obtained, ROC curves were generated for all the features.[21] ROC curve signifies the false positive rate (x-axis) versus the true positive rate (y-axis). It illustrates the performance of a binary system. The ROC curve for the various features were obtained (Fig 2). These curve depicts , to what extent the test is accurate. Also, it shows the trade-off between the specificity and the sensitivity values.

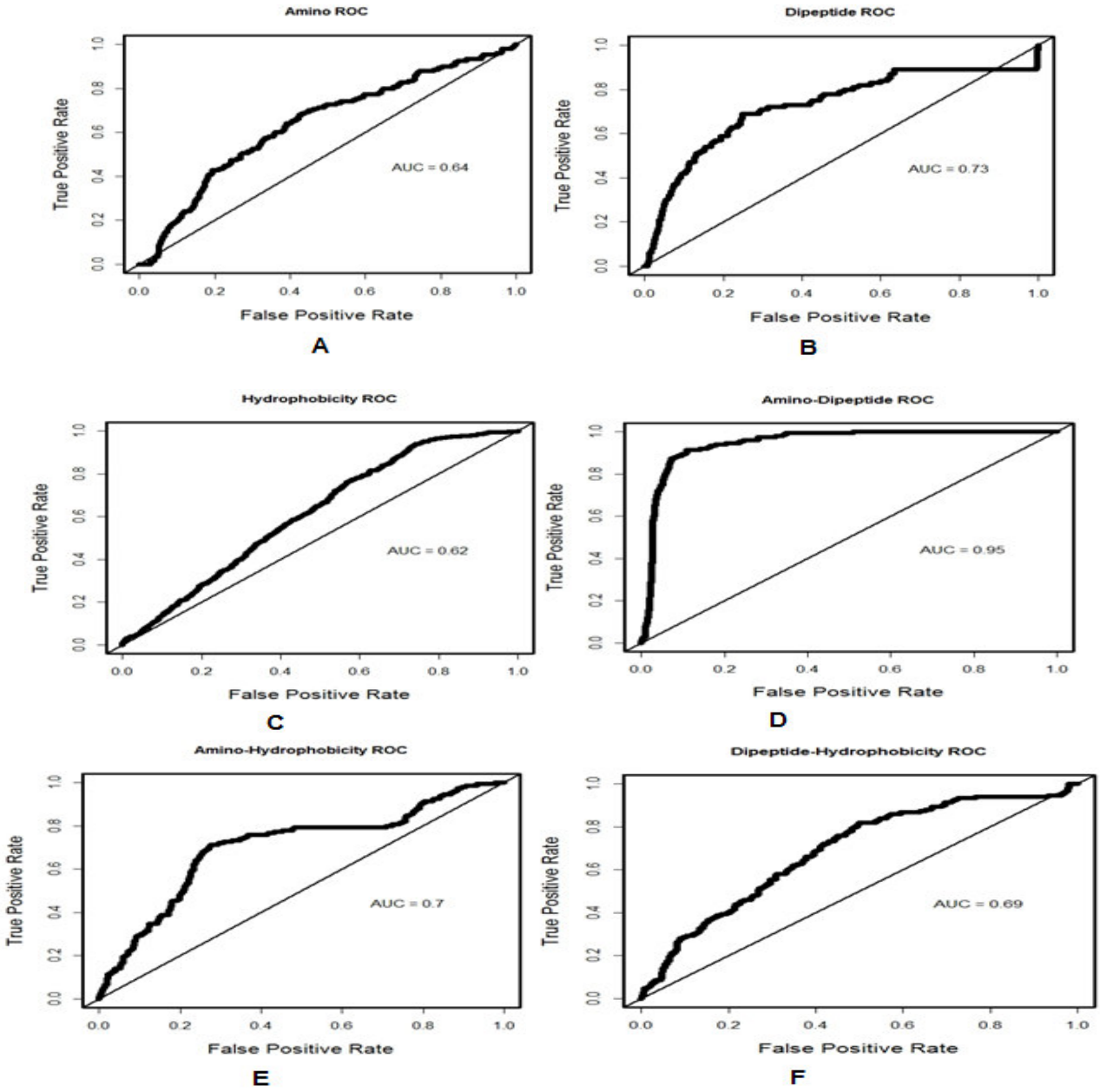

Figure 2

ROC curves obtained for different classes and the hybrid combinations. (A) Amino acid ROC curve. (B)Dipeptide ROC curve. (C)Hydrophobicity ROC curve. (D)Amino acid-Dipeptide ROC curve. (E)Amino acid- Hydrophobicity ROC curve. (F)Dipeptide – Hydrophobicity ROC curve.

From the ROC curves, it can been seen that the ROC curves obtained for almost all the features is closer to the diagonal line. While the one obtained from the AAC-DPC hybrid is shown to have a curve that goes away from the mid-line. When the curve deviates away from the midline, and is closer to the left-hand border, the test is considered to be more accurate. AUC value is also determined using the ROC curve. The Area Under Curve (AUC) value is an evaluation metric parameter that can distinguish between two diagnostic groups. It is observed that, greater the AUC value (be close to 1), the accuracy is said to be greater. The AUC values obtained are being tabulated in Table 4. The AUC value again for the hybrid combination of AAC and DPC is more close to 1 when compared to the other features. From Table2, it could be seen that this combination provides better accuracy when compared to other features.

Table 4

The AUC values for the various features

Variable	AUC
Amino acid	0.64
Dipeptide	0.73
Hydrophobicity	0.62
Amino acid + Dipeptide	**0.95**
Amino acid + Hydrophobicity	0.70
Dipeptide + Hydrophobicity	0.69

CONCLUSION

Polyanion binding proteins play a very crucial role in various biological processes. It has been the basis of various neurodegenerative diseases. Our work deals with the classification of various polyanion binding proteins into different classes. The classification was performed using the program SVM^{light}. Since SVM^{light} s fast optimization algorithm, users can define a number of parameters, kernel functions (radial basis function (RBF) or a polynomial kernel). A classifier module was developed and tested across datasets. The results were further cross-validated using tenfold cross validation technique. Better accuracy was observed and analysis of the ROC curves proved that the classifier designed was effective in classifying the data sets. The other parameters like specificity, sensitivity and MCC values showed that the hybrid feature (AAC+DPC) could effectively help in classification of the dataset.

ACKNOWLEDGEMENTS

This research work is supported by TRR research grant,SASTRA University, Tamil Nadu, India.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Saiz L.,Vilar JM.,Protein-protein/DNA interaction networks: versatile macromolecular structures for the control of gene expression ,IET Syst Biol. 2008; 2(5) ,247-55
2. Jones LS.,Yazzie B.,Middaugh CR.,Polyanions and the proteome, Mol Cell Proteomics. 2004 ;3(8) 746-69
3. Salamat-Miller N.,Fang J.,Seidel CW.,Smalter AM.,Assenov Y.,Albrecht M.,Middaugh CR., A network-based analysis of polyanion-binding proteins utilizing yeast protein arrays,Mol Cell Proteomics. 2006; 5(12) ,2263-78
4. Fang J.,Dong Y.,Salamat-Miller N.,Middaugh CR.,DB-PABP: a database of polyanion-binding proteins, Nucleic Acids Res. 2008; 36,D303-6
5. VAPNIK V., Statistical Learning Theory, Wiley, New York 1998;768
6. Thorsten Joachims.,Text categorization with Support Vector Machines: Learning with many relevant features ,Machine Learning: ECML-98 Lecture Notes in Computer Science. 1998; 1398, 137-142
7. Simon Tong.,Daphne Koller.,Support Vector Machine Active Learning with Applications to Text Classification, Journal of Machine Learning Research. 2001;2,45-66
8. Joachims T.,Text Categorization with Support Vector Machines: Learning with Many Relevant Features ,Springer. 1998 :1398
9. Brown MPS.,Grundy WN.,Lin D.,Cristianini N.,Sugnet C.,Ares JM., Haussler D., Knowledge-based Analysis of Microarray Gene Expression Data by using Support Vector Machines, Proc. Natl.Acad. Sci. 2000; 97,262-267.
10. Ding CHQ.,Dubchak I.,Multi-class Protein Fold Recognition Using Support Vector Machines,Bioinformatics. 2001; 17(4), 349-358
11. Joachims T.,Making large-Scale SVM Learning Practical.,Adv Kernel Methods Support Learn. 1999; 169-184
12. Leinonen R., Diez FG., Binns D., Fleischmann W., Lopez R., Apweiler R., UniProt archive, Bioinformatics. 2004; 20, 3236–3237
13. Ying Huang.,Beifang Niu.,Ying Gao.,Limin Fu.,Weizhong Li.,CD-HIT Suite: a web server for clustering and comparing biological sequences,Bioinformatics. 2010; 26, 680-682
14. Sasisekharan R.,Venkataraman G.,Heparin and heparan sulfate: biosynthesis, structure and function ,Curr Opin Chem Biol. 2000;4(6) ,626-31
15. Baglin T., Barrowcliffe T. W., Cohen A.,Greaves M.,Guidelines on the use and monitoring of heparin,British Journal of Haematology. 2006; 133, 19-34
16. Smith SA.,Mullin NP.,Parkinson J.,Shchelkunov SN.,Totmenin AV.,Loparev VN., Srisatjaluk R., Reynolds DN., Keeling KL., Justus DE., Barlow PN., Kotwal GJ.,Conserved Surface-Exposed K/R-X-K/R Motifs and Net Positive Charge on Poxvirus Complement Control Proteins Serve as Putative Heparin Binding Sites and Contribute to Inhibition of Molecular Interactions with Human Endothelial Cells: a Novel Mechanism for Evasion of Host Defense, Journal Of Virology. 2000; 74(12), 5659–5666
17. Lakshminarasimhan Damodharan.,Vasantha Pattabhi.,Hydropathy analysis to correlate structure and function of proteins, Biochemical and Biophysical Research Communications. 2004; 323 ,996–1002
18. Ryan Rifkin.,Aldebaro Klautau.,In Defense of One-Vs-All Classification ,The Journal of Machine Learning Research. 2004; 5,101-141
19. Morik K., Brockhausen P., Joachims T.,Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring, Proc. 16th Int'l Conf., on Machine Learning (ICML-99), 1999.
20. Shao-Wu Zhang.,Quan Pan.,Hong-Cai Zhang.,Yun-Long Zhang.,Hai-Yu Wang., Classification of protein quaternary structure with support vector machine, People's Republic of China. 2003; 19(18), 2390–2396
21. Christopher M F., Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests ,Clin Biochem Rev. 2008; 29, S83–S87.